Geometry-Aware and Semi-Supervised Ensembles for Multimodal Mitral Valve Segmentation

Matej Gazda and David Paik

Laza Medical, Campbell, CA 95008, USA
matej@lazamedical.com, david@lazamedical.com

Abstract. Mitral leaflets are thin, folded sheets that coapt along a curved line, and the MVAA 2026 challenge requires the same valve to be recovered from cardiac CT, 3-D transesophageal echocardiography (TEE) and surgical video. We describe the three modality-specific systems in our submitted container. For CT we make the leaflet's thin-sheet geometry explicit instead of leaving it to be learned from intensity: the network is given a 47-channel bank of multi-scale Gaussian derivatives, Hessian entries and sheetness features, and is trained with a shape-aware CarveMix whose donor region follows a level set of the label's distance transform and is mixed on raw Hounsfield values before the descriptor is computed. On the hidden test this scores 0.8494 DSC, 4.96 mm HD_{100} and 0.404 mm ASD. For TEE, a cleanup that removes only distant components and then trims filaments with a geodesically restricted opening reaches 0.8583 DSC, 9.17 mm HD_{100} and 0.497 mm ASD. For surgical video, three DINOv3 ViT-L/DPT profiles combining GIN augmentation and an auxiliary valve/ventricle head, trained with Mean Teacher on 1,379 unlabelled frames drawn from other surgeries, reach 0.8386 DSC, 170.5 px HD_{100} and 35.8 px ASD. We additionally report a negative methodological finding: at the effect sizes separating candidate systems, neither cross-validation nor the 30/20/48-case public validation set ordered them consistently with the hidden test, and one Task 1 HD_{100} comparison inverted in sign.

Keywords: Mitral valve · multimodal segmentation · nnU-Net · semi-supervised learning · surgical video

1 Introduction

The mitral valve seals the left atrioventricular orifice with a saddle-shaped annulus and two thin, coapting leaflets that meet along a curved coaptation line. The leaflets are soft, mobile, and only a few voxels thick in volumetric images, and their shape carries much of the information needed when planning repair for mitral regurgitation. Quantifying this anatomy is therefore useful but difficult, and harder still when the same structure must be recovered from different acquisitions, which is what the MVAA 2026 challenge poses as three segmentation tasks over one valve [7]: in cardiac CT the leaflet is a small sheet inside a high-resolution volume; in 3-D TEE, signal dropoff and speckle obscure the boundary between

adjacent leaflets; in surgical video the valve is a soft-tissue region on a moving scene with instruments, blood, and frames where the valve is absent from the image.

Existing mitral pipelines commonly derive an explicit leaflet surface from a voxel segmentation [6,3], whose correspondence and topology assumptions sit awkwardly on a sheet that folds and coapts. Dense masks do not inherently constrain morphology to anatomically realistic configurations, but they are what the challenge scores, so we keep dense segmentation as the output and add modality-specific priors instead: a geometric prior for the volumetric tasks, describing the leaflet's multi-scale sheet structure alongside intensity, and a representational one for video, where six annotated surgeries cannot train an encoder from scratch.

We contribute (i) a geometry-aware CT input, a 47-channel Hessian and sheetness descriptor trained with a shape-aware CarveMix [12] variant applied to raw intensities before the descriptor is computed; (ii) a conservative TEE post-processing cleanup, deleting only distant components and filaments, that cuts HD_{100} by 1.74 mm without reducing Dice; (iii) a semi-supervised video ensemble of three DINOv3 ViT-L/DPT profiles combining GIN augmentation and an auxiliary valve/ventricle head; and (iv) evidence that at these cohort sizes internal cross-validation and the public validation set did not order candidate systems consistently with the hidden test, including a sign inversion on Task 1 HD_{100}.

2 Data and evaluation

Task 1 provides 27 labelled and 1,040 unlabelled CT volumes plus 30 public-validation cases; Task 2, 105 labelled and 20 public 3-D TEE acquisitions with anterior and posterior leaflet classes; Task 3, 179 quality-controlled frames from six surgical videos (117 valve-present), 1,379 unlabelled frames from 46 further videos and 48 public frames [7]. A separate test set is withheld by the organisers and never released; they score it themselves by running a submitted Docker image. We call it the hidden test set. The official evaluator reports the Dice similarity coefficient (DSC), the Hausdorff distance between the two mask surfaces (HD_{100}) and the average symmetric surface distance (ASD) per task as nine separate columns, with no composite, so we report all three metrics for every decision; distances are symmetric and bidirectional, in millimetres for CT and TEE and pixels for video, and HD_{100} is the maximum rather than a percentile.

3 Methods

Training. One CT model trains in 9.9 h on a single NVIDIA H200: 250 epochs of 250 iterations at batch size 2 and 192^3 voxels with 47 input channels, 142 s per epoch on an idle GPU, and 143 s with CarveMix, i.e. budget-matched. One TEE model follows the same nnU-Net schedule for 250 epochs. One video model costs

105 s per epoch and reaches its best checkpoint at epoch 7–13, i.e. 12–23 min, well inside a 40-epoch budget with early stopping.

Inference. The submission is a single Docker image running all three tasks offline with fixed weights, thresholds and post-processing; nothing is learned or re-fitted at test time. It processes the full public validation set (30 CT, 20 TEE, 48 frames) end to end in 1,698 s on one RTX 3090 Ti; the organisers ran the same image against the hidden test on a 32 GB V100.

Validation protocol. Tasks 1 and 2 use case-disjoint five-fold cross-validation, the nnU-Net default, and Task 3 uses six leave-one-video-out (LOVO) folds, one per annotated surgery, so that count is set by the data; no frame from an evaluated surgery is seen in training. Internal Tasks 1–2 numbers score each case by the fold that did not train on it, and internal Task 3 numbers are pooled over the six held-out videos under one common refit protocol. The two Task 3 settings are not identical: internally each frame receives one checkpoint per profile, the one that held its video out, whereas the container averages all six folds of every profile, so the deployed ensemble is six times larger than the one we score internally. A per-profile logit bias, fitted on the pooled out-of-fold predictions, puts the three profiles on a common operating point before averaging. All operating points and post-processing parameters were selected on those pooled out-of-fold predictions, never on the public validation set, so they are post-hoc rather than nested fits whose optimism we quantify below.

3.1 Task 1: geometry-aware CT segmentation

A thin sheet has a characteristic second-order signature: at a scale matched to its thickness the image Hessian has one large eigenvalue along the sheet normal and two small ones in the sheet plane, measured by convolving the image with the second-order partial derivatives of a 3-D Gaussian at several scales. Frangi et al. build these eigenvalue features and then suppress the plate-like case to isolate vessels [2], leaving unused the planar case a leaflet exhibits; the eigenvalues are rotation invariant, whereas a raw derivative bank is only equivariant, since rotating the image remixes its channels. We supply this structure explicitly, in closed form, rather than leaving it to be learned in approximation with potentially uneven behaviour across scales and orientations.

All CT models are large residual-encoder nnU-Nets [4,5] with 192^3 patches at 0.3 mm isotropic resampling, CT normalisation, Dice and cross-entropy loss, and 250 epochs of 250 iterations in five folds. The input has 47 channels: normalised CT (1), a multi-scale Gaussian-derivative bank (40: smoothed intensity, the three first derivatives and the six independent Hessian entries at $\sigma = 0.45, 0.9, 1.8, 3.6$ mm), and six sheetness features (windowed CT, the maximum multi-scale sheetness response, the scale attaining it, and the three signed Hessian eigenvalues sorted by magnitude), computed by a compiled separable-Gaussian CUDA kernel inside the training loop. The bank is taken on the z-scored volume, the sheetness features on a second bank built from an HU-windowed $[0, 700]$ copy. Sheetness is evaluated in closed form from the σ^2-normalised eigenvalues $|\lambda_1| \leq |\lambda_2| \leq |\lambda_3|$, in Frangi's form but sign-agnostic and rewarding rather than suppressing the plate case, as

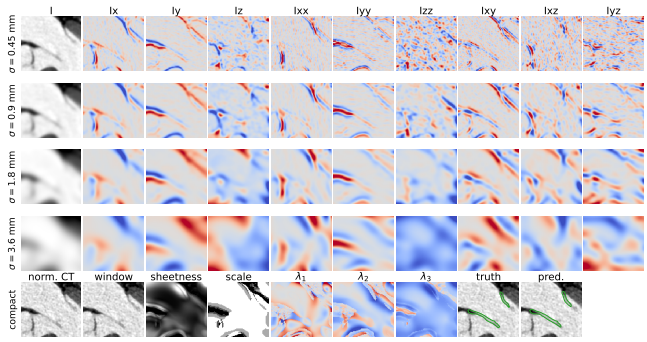

Fig. 1. The 47-channel input the CT network receives, shown for one labelled case and cropped around the leaflet. Rows one to four are the Gaussian/Hessian bank at $\sigma = 0.45, 0.9, 1.8, 3.6$ mm: smoothed intensity, first derivatives, and the six independent second derivatives (red positive, blue negative). The bottom row holds the six compact channels, the ground truth and the prediction (green contour). Fine scales resolve the leaflet edge, coarse scales capture the sheet, and sheetness responds along the leaflet rather than at generic intensity edges.

$\exp\big[-(\lambda_2/\lambda_3)^2/2\alpha^2\big]\,\big[1-\exp(-S^2/2\beta^2)\big]$ with $\alpha = 0.5$, $\beta = 0.1$, $S = \|\lambda\|_2$ and the denominator clamped away from zero, maximised over the four scales. The network could in principle recover this from the raw Hessian entries it already receives, so the ablation row measures whether supplying the summary helps, not whether the information is new. The whole bank is computed after augmentation, so derivatives stay aligned with the image the network is given (Fig. 1).

The submitted family adds shape-aware copy-paste [12]: for a donor case we take the Chebyshev distance transform of its leaflet label and the region enclosed by a level set at $\lambda \sim \mathcal{U}\{-2, \ldots, +8\}$ voxels (0.3 mm each), so the carved region hugs the annotation instead of cutting a straight artificial edge through a 1 mm sheet, with per-sample mix probability 0.5. The mix is applied to the raw Hounsfield tensor *before* the bank is computed, because the descriptor is non-linear in intensity and mixing afterwards would splice together channel statistics no real scan produces. At inference the five folds are averaged with mirrored test-time augmentation (10 forward passes per case), thresholded at 0.45, and reduced to the largest 26-connected component.

3.2 Task 2: 3-D TEE with an anisotropic residual U-Net

TEE volumes are resampled to the median dataset spacing and segmented by a medium residual-encoder nnU-Net trained for 250 epochs in five case-disjoint folds with multiclass Dice and cross-entropy; probabilities are averaged over folds before the argmax. Mirroring is restricted to the two axes along which leaflet identity is preserved, since a geometry audit found genuine x-orientation reversals in 14 of 105 acquisitions, so "posterior is on one side" is not an invariant of the array coordinates.

Post-processing, called *cleanup* below, runs per class on 26-connected components, only ever deletes predicted voxels, and is deliberately conservative. Besides the largest component, a component survives if its physical volume reaches 200 mm^3, or if it holds at least 5% of the largest component's volume and lies within 8 mm of it in minimum Euclidean distance under physical spacing. Everything else is removed. Thin protrusions survive that filter because they are attached, so a geodesically restricted opening follows: erode by one voxel with a 6-connected footprint, dilate back to a core, and delete only voxels outside the core that lie more than 1.2 mm from it. A filament thinner than the footprint falls entirely outside the core and dies at its tip, whereas the sheet keeps a shallow rim and is preserved exactly, so Dice does not move. Any component removed in full but larger than 30 mm^3 is restored, which protects genuine thin scallops.

3.3 Task 3: surgical video with a DINOv3 ensemble, Mean Teacher and GIN

With 179 labelled frames from six surgeries the encoder is pretrained rather than trained, and the unlabelled pool supplies consistency targets [11]. The submitted system, *ginTrio*, averages three profiles with equal weight, each contributing its six LOVO checkpoints, so 18 in total. Every profile is a DINOv3 ViT-L/16 encoder, pretrained on the 1.7B-image LVD-1689M web corpus rather than on the satellite variant, with a DPT dense decoder [10,9], which reassembles tokens from several transformer stages and contains no batch normalisation. Both parts are fine-tuned, the encoder at learning rate 1.5×10^{-5} and the decoder at 2×10^{-4}, at effective batch 8 throughout, with a Dice plus focal objective (0.7/0.3, $\gamma = 2.0$, $\alpha = 0.55$); one profile runs at 704×1248 and the other two at 640×1152. All three profiles also train on 600 synthetic occlusion frames that we generate, at loss weight 0.5. Donors are annotated suture, needle and pledget components cut from *other* surgeries and pasted into a recipient frame, then repainted inside the paste mask by a masked GAN renderer trained on real occluders, because plain alpha compositing leaves seams the network exploits. Placement follows three profiles: half cross the valve's principal axis and are rejected unless they raise its connected-component count, a quarter sit on background with the label unchanged so the set carries no shrink bias, and a quarter put cross-video ventricular tissue in the annular ring as a hard negative. Occluded valve area is erased from that sample's label, a median 16.8% of the recipient frame's annotated valve area on the crossing samples. Erasing matches the organisers' convention rather than contradicting it: their masks stop at the visible surface, so occluded valve pixels count as background when scored. Measuring the concavities recovered by morphologically closing the valve mask, occluder classes fill 71% of that filled-in area against 17% of the frame, a 4.1× enrichment, and suture, needle and pledget dominate it. The half loss weight is load-bearing, since trusting these frames fully measured worse than omitting them.

Mean Teacher supervises the student on the 1,379 unlabelled frames with EMA decay 0.996, an eight-epoch warm-up and a 16-epoch ramp to weight 0.7, accepting a pseudo-target only above 0.85 foreground or below 0.05 with a

64-pixel minimum area. The warm-up length is consequential: the pool comes from 46 *other* surgeries, so engaging the consistency term early measured −0.0275 DSC, and the shipped checkpoints are selected at epochs 7–13, where the term still carries a small fraction of its target weight.

The profiles are the plain configuration; the same plus an auxiliary three-way valve/ventricle head on separate output channels at weight 0.50, excluding a three-pixel seam at the ambiguous interface, with the scored binary channel receiving zero gradient from the head (gradient-checked); and a third carrying that head, the higher resolution and GIN augmentation [8], which randomises appearance with randomly initialised non-linear filters: four stacked random convolutions (3×3 then 1×1) with LeakyReLU between them, renormalised to the source per-channel statistics and blended with $\alpha \sim \mathcal{U}(0, 0.5)$ at probability 0.5, all stride 1 with padding so no pixel moves and the mask stays valid. The pack is therefore a deliberately heterogeneous ensemble, not a one-variable ablation: the head is isolated by the first two profiles and GIN by a separate control matched to the third in every other flag. At inference each checkpoint predicts the original and horizontally flipped frame at native resolution, a cheap choice that a 29-augmentation echocardiography study also finds among the few transformations improving cross-dataset transfer, an effect they attribute to spurious associations being broken so the model leans on image evidence instead [1], a per-profile logit bias puts the profiles on a common operating point, the 18 maps are averaged, and the threshold comes from a dense global sweep on pooled out-of-fold predictions.

4 Results

Table 1 gives the submitted container's scored results together with the ablations behind them. The two organiser-scored splits are different cohorts, so their columns should be read separately: the submitted system scores 0.246 mm ASD on public validation and 0.404 mm on the hidden test. Configurations are comparable only within one column.

Task 1. Explicit geometry and CarveMix each helped. The four Task 1 rows of Table 1 each add one component to the row above, so they show what each addition is worth in this order, not what each component is worth alone. Intensity alone gives 0.8555 DSC. The four-scale 40-channel bank buys little Dice but 0.29 mm of HD_{100}. Six sheetness channels on top of that unchanged bank add 0.0034 DSC and give back 0.09 mm, a Dice-for-boundary trade rather than a uniform gain. CarveMix raises Dice by 0.0053 and cuts ASD by 0.017 mm over its own control, and both survive a paired test across the 27 cases: DSC +0.0053 (95% bootstrap CI +0.0008 to +0.0095, Wilcoxon $p = 0.006$, 19/27 cases), ASD −0.017 mm (CI −0.029 to −0.006, $p < 0.001$, 22/27). Its HD_{100} gain of 0.22 mm does not (CI −0.66 to +0.07, $p = 0.50$, 14/27).

Task 2. Cleanup buys boundary accuracy at Dice parity: Table 1 shows HD_{100} falling by 1.74 mm out of fold while Dice moves by +0.0003. Changing the target or the loss did not help. Separate anterior and posterior decoders, a Hausdorff-distance loss, a TopK loss, a clDice loss aimed at thin structure

Table 1. Every system, internal cross-validation against both organiser-scored splits; checkpoint counts in parentheses. Task 1 internal is out-of-fold on 27 cases with pooled post-hoc threshold calibration per arm, Task 2 out-of-fold on 105 cases, Task 3 pooled six-fold LOVO. Distances are mm for Tasks 1–2 and pixels for Task 3. Bold marks the best value in each column; dashes mark arms never scored on that split.

	Internal (cross-validated)			Public validation			Hidden test		
Configuration	DSC↑	HD↓	ASD↓	DSC↑	HD↓	ASD↓	DSC↑	HD↓	ASD↓
Task 1: CT, incremental descriptor at 192^3 *(5 folds)*									
intensity only	0.8555	4.34	0.260	–	–	–	–	–	–
+ four-scale bank (41 ch)	0.8569	4.05	0.255	–	–	–	–	–	–
+ sheetness (47 ch)	0.8603	4.14	0.254	0.8627	**4.18**	0.257	–	–	–
+ CarveMix (**submitted**)	**0.8656**	**3.92**	**0.237**	**0.8653**	4.25	**0.246**	0.8494	4.96	0.404
Task 2: 3-D TEE (5 folds)									
ResEnc-M, raw	0.8360	12.23	0.684	–	–	–	–	–	–
+ filter + trim (**subm.**)	0.8363	10.49	0.661	0.8450	10.20	0.603	0.8583	9.17	0.497
Task 3: surgical video (6 leave-one-video-out folds)									
ConvNeXt DINOv3 ens. (72)	0.7970	95.5	18.3	0.8174	71.3	12.6	–	–	–
ViT-L pair (12)	0.8198	85.8	14.7	0.8222	**66.8**	11.6	–	–	–
GIN solo (6)	0.8229	82.3	14.2	0.8219	68.7	**11.4**	–	–	–
ginTrio (18, **submitted**)	**0.8263**	**81.8**	**13.6**	**0.8243**	67.7	11.5	0.8386	170.5	35.8

(0.8422 against 0.8457 Dice on a common fold) and unrestricted mirroring were all neutral or negative.

Task 3. Encoder choice dominated. ViT-L/DPT improved all three metrics over the ConvNeXt-based DINOv3 ensemble, and adding GIN produced the submitted scores. Of the domain-generalisation augmentations only GIN helped, and its mechanism is measurable: a per-channel affine fit leaves a residual of 0.1246 on GIN output, whereas one linear random convolution is invertible by such an affine, which the network learns to exploit. Ensemble size does not explain the result, since on public validation Table 1 puts GIN solo within 0.0024 DSC of ginTrio on six checkpoints rather than eighteen, with better ASD. Backbone and decoder swaps failed, as did boundary losses, a residual refiner, component filtering, CutMix, pseudo-labels and resolutions of 640–1408. The dominant error is anatomical: 28.3% of false-positive pixels fall on annotated ventricular tissue against $\approx 3.5\%$ on all twelve instrument classes combined, which is what the auxiliary head targets.

Agreement between evaluation splits. The two internal splits did not order candidates as the hidden test did. The CarveMix swap measured +0.0026 DSC, +0.07 mm HD_{100} (worse) and −0.011 mm ASD on the 30-case validation set, against +0.0083, −1.03 mm (better) and −0.122 mm on the hidden test. HD_{100} inverted in sign. For an extreme-order statistic on a small cohort this is expected: a bootstrap over our 27 internal cases puts a ±0.42 mm band on mean HD_{100}, an order of magnitude wider than the differences we were trying to resolve. Task 3 shows the same failure on Dice, where pooled out-of-fold scores inverted against the board at least six times; the clearest is a four-profile pack with the best internal Dice we recorded (0.8313) that scored lowest of its cohort (0.8242). Two magnitudes bound their resolution: the 48-frame public set has a Dice noise floor near 0.014, and seed-only replicas spanned 0.0072 Dice.

5 Discussion and conclusion

The components that survived were those aimed at a measured error: the thin-sheet descriptor and shape-aware mix for CT, distance-aware component removal and a geodesic trim for TEE, and a pretrained encoder with intensity-domain augmentation for video. The evaluation itself was the harder problem. On cohorts this small our internal scores did not order candidates as the blind sets did, and the disagreement ran both ways: internal gains shrank or reversed on the hidden test, and a configuration that lost internally won on the public board. A change supported only by a small HD_{100} difference on a small held-out set should therefore be treated as unresolved at this sample size rather than as an improvement: the paired test that confirms our CarveMix DSC and ASD gains leaves the HD_{100} gain of the same system indistinguishable from no change. Our own limitations follow from the same fact. The parts of the CT descriptor are never separated from each other, its thresholds are fitted on pooled rather than cross-fitted predictions, Task 3 rests on six videos, and our internal estimates cannot show that the shipped configuration is the best we built, only the best we scored.

References

1. Elyasi, S., Adibzadeh, S., Dadashi Serej, N., Zolgharni, M.: EAGT: Echocardiography augmentation for generalisability and transferability. arXiv preprint arXiv:2605.16427 (2026)
2. Frangi, A.F., Niessen, W.J., Vincken, K.L., Viergever, M.A.: Multiscale vessel enhancement filtering. In: MICCAI. LNCS, vol. 1496, pp. 130–137. Springer (1998)
3. Huynh, P.K., et al.: 3-D TEE mitral valve segmentation and mesh reconstruction with real-time quality assurance. In: IEEE BHI. pp. 1–7 (2025)
4. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (2021)
5. Isensee, F., et al.: nnU-Net revisited: A call for rigorous validation in 3D medical image segmentation. In: MICCAI. LNCS, vol. 15009, pp. 488–498. Springer (2024)
6. Munafò, R., et al.: A deep learning-based fully automated pipeline for regurgitant mitral valve anatomy analysis from 3D echocardiography. IEEE Access **12**, 5295–5308 (2024)
7. MVAA Challenge Organizers: Mitral valve anatomy analysis using multimodal imaging data (MVAA 2026). CodaBench competition documentation (2026)
8. Ouyang, C., et al.: Causality-inspired single-source domain generalization for medical image segmentation. IEEE Trans. Med. Imag. **42**(4), 1095–1106 (2023)
9. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: ICCV. pp. 12179–12188 (2021)
10. Siméoni, O., et al.: DINOv3 (2025), `https://arxiv.org/abs/2508.10104`
11. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: NeurIPS. pp. 1195–1204 (2017)
12. Zhang, X., et al.: CarveMix: A simple data augmentation method for brain lesion segmentation. In: MICCAI. LNCS, vol. 12901, pp. 196–205 (2021)